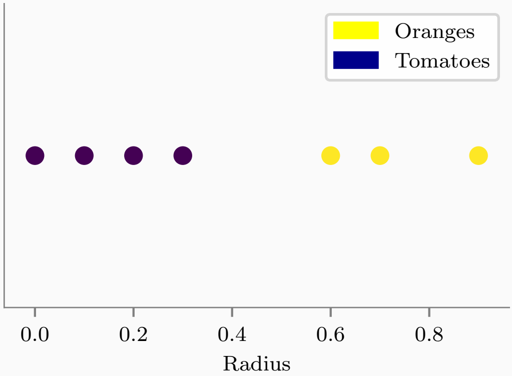
Logistic Regression

Nipun Batra

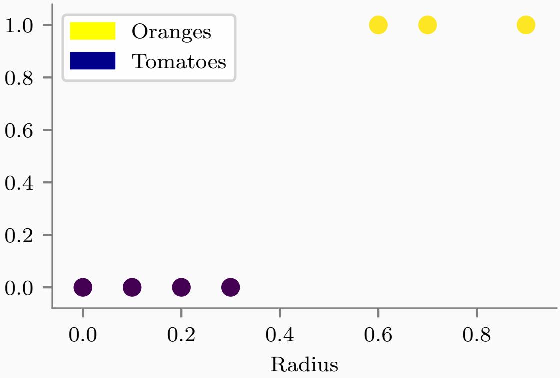
February 18, 2020

IIT Gandhinagar

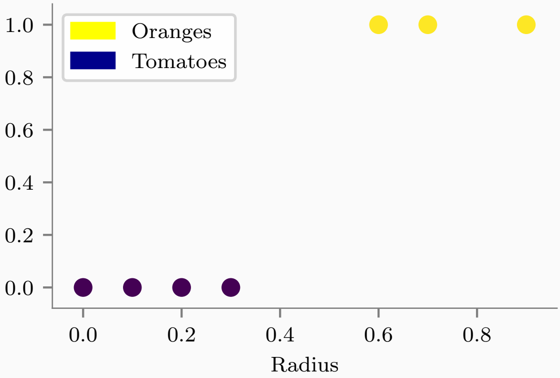
Classification Technique



Classification Technique

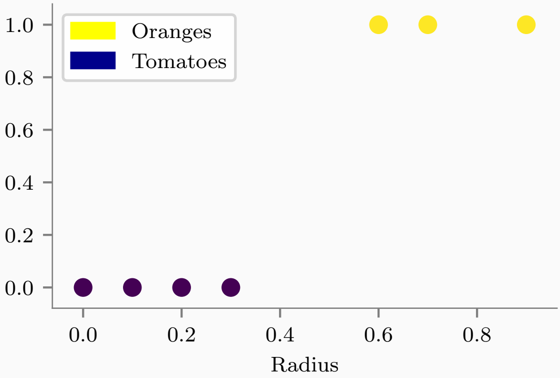


Classification Technique



Aim: Probability(Tomatoes | Radius) ? or

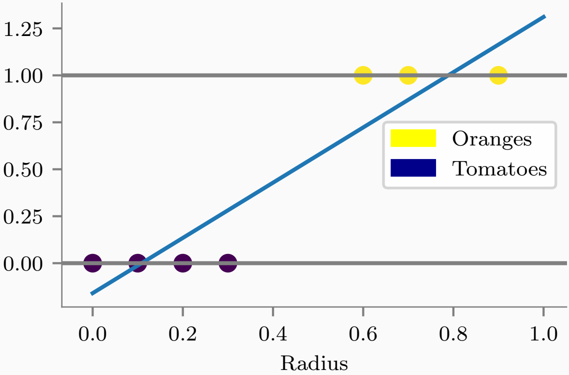
Classification Technique



Aim: Probability(Tomatoes | Radius) ? or

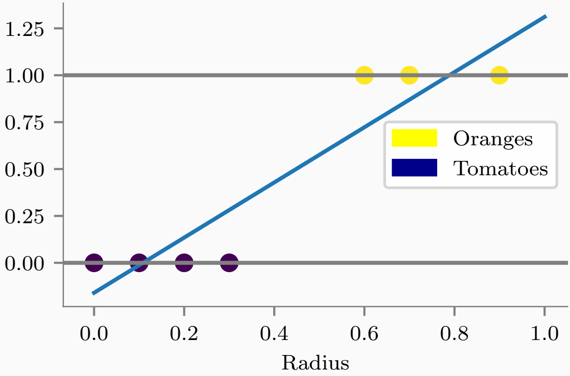
More generally, $P(y = 1|X = x)$?

Idea: Use Linear Regression



$$P(X = Orange|Radius) = \theta_0 + \theta_1 \times Radius$$

Idea: Use Linear Regression



$$P(X = Orange|Radius) = \theta_0 + \theta_1 \times Radius$$

Generally,

$$P(y = 1|x) = X\theta$$

Idea: Use Linear Regression

Prediction:

If $\theta_0 + \theta_1 \times Radius > 0.5 \rightarrow$ Orange

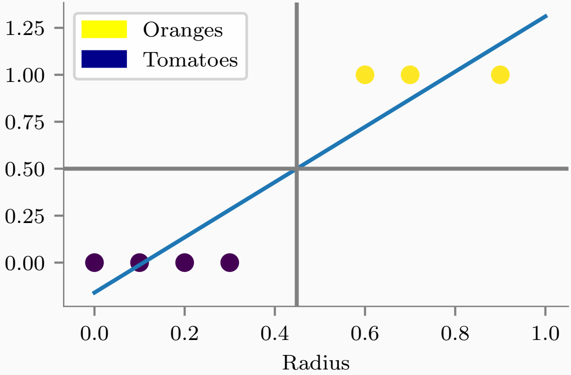
Else $\rightarrow$ Tomato

Problem:

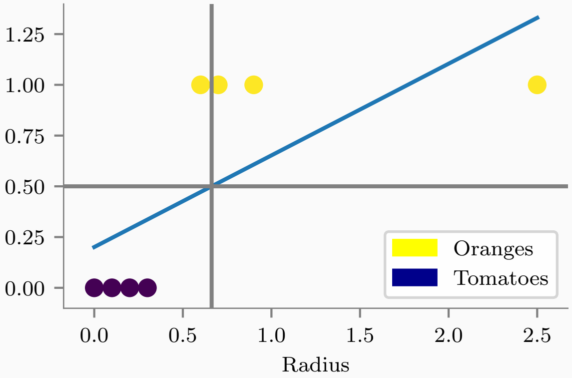
Range of $X\theta$ is $(-\infty, \infty)$

But $P(y = 1|\ldots) \in [0, 1]$

Idea: Use Linear Regression

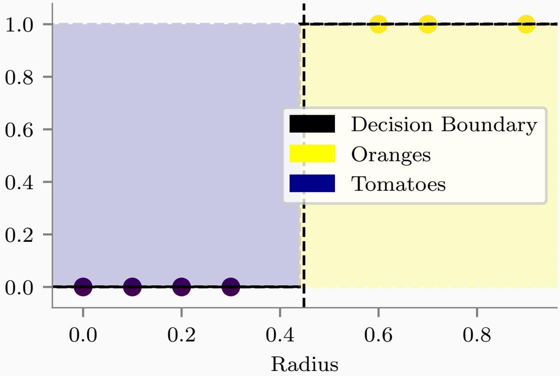


Idea: Use Linear Regression



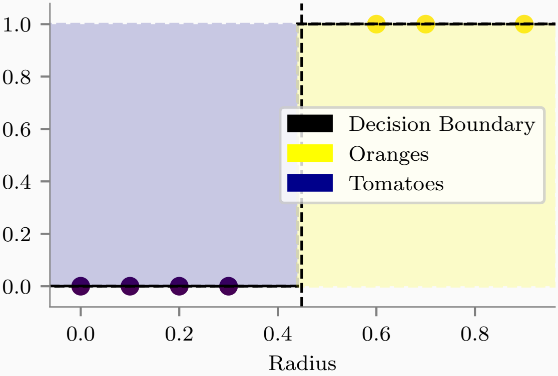
Linear regression for classification gives a poor prediction!

Ideal boundary



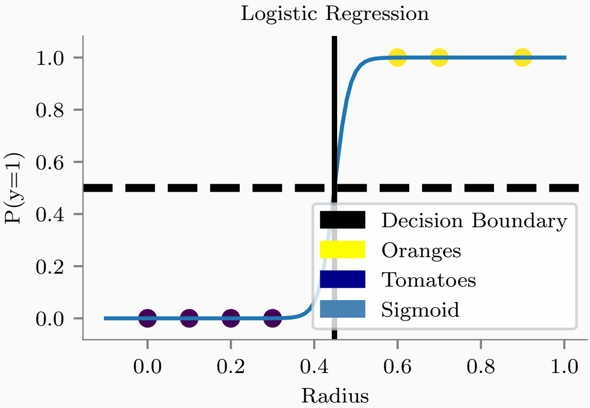
- Have a decision function similar to the above (but not so sharp and discontinuous)

Ideal boundary



- Have a decision function similar to the above (but not so sharp and discontinuous)
- Aim: use linear regression still!

Idea: Use Linear Regression



Question. Can we still use Linear Regression?

Answer. Yes! Transform $\hat{y} \to [0, 1]$

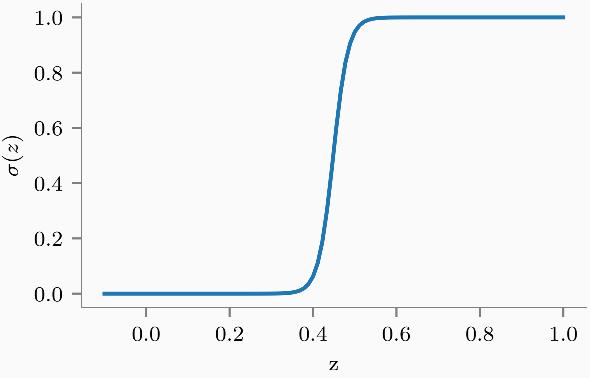
Logistic / Sigmoid Function

$\hat{y} \in (-\infty, \infty)$

ϕ = Sigmoid / Logistic Function (σ)

$\phi(\hat{y}) \in [0, 1]$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Logistic / Sigmoid Function

$z \to \infty$

Logistic / Sigmoid Function

$z \rightarrow \infty$

$\sigma(z) \rightarrow 1$

Logistic / Sigmoid Function

$z \to \infty$

$\sigma(z) \to 1$

$z \to -\infty$

Logistic / Sigmoid Function

$z \to \infty$

$\sigma(z) \to 1$

$z \to -\infty$

$\sigma(z) \to 0$

Logistic / Sigmoid Function

$z \to \infty$

$\sigma(z) \to 1$

$z \to -\infty$

$\sigma(z) \to 0$

$z = 0$

Logistic / Sigmoid Function

$z \to \infty$

$\sigma(z) \to 1$

$z \to -\infty$

$\sigma(z) \to 0$

$z = 0$

$\sigma(z) = 0.5$

Logistic / Sigmoid Function

Question. Could you use some other transformation (ϕ) of $\hat{y}$ s.t.

$$\phi(\hat{y}) \in [0, 1]$$

Yes! But Logistic Regression works.

Logistic / Sigmoid Function

$$P(y=1|X) = \sigma(X\theta) = \frac{1}{1+e^{-X\theta}}$$

Q. Write $X\theta$ in a more convenient form (as $P(y=1|X)$, $P(y=0|X)$)

Logistic / Sigmoid Function

$$P(y = 1|X) = \sigma(X\theta) = \frac{1}{1 + e^{-X\theta}}$$

Q. Write $X\theta$ in a more convenient form (as $P(y = 1|X)$, $P(y = 0|X)$)

Logistic / Sigmoid Function

$$P(y = 1|X) = \sigma(X\theta) = \frac{1}{1 + e^{-X\theta}}$$

Q. Write $X\theta$ in a more convenient form (as $P(y = 1|X)$, $P(y = 0|X)$)

$$P(y = 0|X) = 1 - P(y = 1|X) = 1 - \frac{1}{1 + e^{-X\theta}} = \frac{e^{-X\theta}}{1 + e^{-X\theta}}$$

Logistic / Sigmoid Function

$$P(y = 1|X) = \sigma(X\theta) = \frac{1}{1 + e^{-X\theta}}$$

Q. Write $X\theta$ in a more convenient form (as $P(y = 1|X)$, $P(y = 0|X)$)

$$P(y = 0|X) = 1 - P(y = 1|X) = 1 - \frac{1}{1 + e^{-X\theta}} = \frac{e^{-X\theta}}{1 + e^{-X\theta}}$$

$$\therefore \frac{P(y = 1|X)}{1 - P(y = 1|X)} = e^{X\theta} \implies X\theta = \log \frac{P(y = 1|X)}{1 - P(y = 1|X)}$$

Odds (Used in betting)

$$\frac{P(win)}{P(loss)}$$

Here,

$$Odds = \frac{P(y=1)}{P(y=0)}$$

$$\boxed{\text{log-odds} = \log \frac{P(y=1)}{P(y=0)} = X\theta}$$

Logistic Regression

Q. What is decision boundary for Logistic Regression?

Q. What is decision boundary for Logistic Regression?

Decision Boundary: $P(y=1|X) = P(y=0|X)$

$$\text{or } \frac{1}{1+e^{-X\theta}} = \frac{e^{-X\theta}}{1+e^{-X\theta}}$$

$$\text{or } e^{X\theta} = 1$$

$$\text{or } X\theta = 0$$

Learning Parameters

Could we use cost function as:

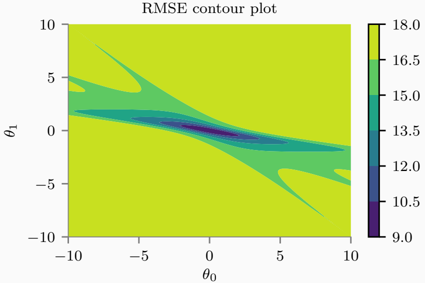
$$J(\theta) = \sum (y_i - \hat{y}_i)^2$$

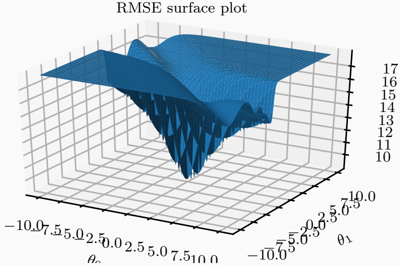
$$\hat{y}_i = \sigma(X\theta)$$

Answer: No (Non-Convex)

(See Jupyter Notebook)

Cost function convexity





Learning Parameters

Likelihood = $P(D|\theta)$

$P(y|X,\theta) = \prod_{i=1}^{n} P(y_i|x_i,\theta)$
where y = 0 or 1

Learning Parameters

Likelihood = $P(D|\theta)$

$$P(y|X,\theta) = \prod_{i=1}^{n} P(y_i|x_i,\theta)$$
$$= \prod_{i=1}^{n} \left\{\frac{1}{1+e^{-x_i^T\theta}}\right\}^{y_i} \left\{1 - \frac{1}{1+e^{-x_i^T\theta}}\right\}^{1-y_i}$$

[Above: Similar to $P(D|\theta)$ for Linear Regression;

Difference Bernoulli instead of Gaussian]

$$-\log P(y|X,\theta) = \text{Negative Log Likelihood}$$
$$= \text{Cost function will be minimising}$$
$$= J(\theta)$$

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: 4/10 = 0.4. But why?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: 4/10 = 0.4. But why?
- Answer 1: Probability defined as a measure of long running frequencies

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: 4/10 = 0.4. But why?
- Answer 1: Probability defined as a measure of long running frequencies
- Answer 2: What is likelihood of seeing the above sequence when the p(Head)=θ?

Aside on Bernoulli Likelihood

- Assume you have a coin and flip it ten times and get (H, H, T, T, T, H, H, T, T, T).
- What is p(H)?
- We might think it to be: 4/10 = 0.4. But why?
- Answer 1: Probability defined as a measure of long running frequencies
- Answer 2: What is likelihood of seeing the above sequence when the p(Head)=θ?
- Idea find MLE estimate for θ

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1 - \theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$
- Log-likelihood = $\mathcal{LL}(\theta) = n_h \log(\theta) + n_t \log(1-\theta)$

Aside on Bernoulli Likelihood

- $p(H) = \theta$ and $p(T) = 1 - \theta$
- What is the PMF for first observation $P(D_1 = x|\theta)$, where x = 0 for Tails and x = 1 for Heads?
- $P(D_1 = x|\theta) = \theta^x(1-\theta)^{(1-x)}$
- Verify the above: if x = 0 (Tails), $P(D_1 = x|\theta) = 1 - \theta$ and if x = 1 (Heads), $P(D_1 = x|\theta) = \theta$
- What is $P(D_1, D_2, ..., D_n|\theta)$?
- $P(D_1, D_2, ..., D_n|\theta) = P(D_1\theta)P(D_2|\theta)...P(D_n|\theta)$
- $P(D_1, D_2, ..., D_n|\theta) = \theta^{n_h}(1-\theta)^{n_t}$
- Log-likelihood = $\mathcal{LL}(\theta) = n_h \log(\theta) + n_t \log(1-\theta)$
- $\frac{\partial \mathcal{LL}(\theta)}{\partial \theta} = 0 \implies \frac{n_h}{\theta} + \frac{n_t}{1-\theta} = 0 \implies \theta_{MLE} = \frac{n_h}{n_h + n_t}$

Learning Parameters

$$J(\theta) = -\log \left\{ \prod_{i=1}^{n} \left\{ \frac{1}{1+e^{-x_i^T\theta}} \right\}^{y_i} \left\{ 1 - \frac{1}{1+e^{-x_i^T\theta}} \right\}^{1-y_i} \right\}$$

$$J(\theta) = -\left\{ \sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1-y_i)\log(1-\sigma_\theta(x_i)) \right\}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{\partial}{\partial \theta_j} \left\{ \sum_{i=1}^{n} y_i log(\sigma_\theta(x_i)) + (1-y_i) log(1-\sigma_\theta(x_i)) \right\}$$

$$= -\sum_{i=1}^{n} \left[y_i \frac{\partial}{\partial \theta_j} \log(\sigma_\theta(x_i)) + (1-y_i) \frac{\partial}{\partial \theta_j} log(1-\sigma_\theta(x_i)) \right]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n} \left[y_i \frac{\partial}{\partial \theta_j} \log(\sigma_\theta(x_i)) + (1 - y_i) \frac{\partial}{\partial \theta_j} log(1 - \sigma_\theta(x_i)) \right]$$

$$= -\sum_{i=1}^{n} \left[\frac{y_i}{\sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} \sigma_\theta(x_i) + \frac{1 - y_i}{1 - \sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} (1 - \sigma_\theta(x_i)) \right] \quad (1)$$

Aside:

$$\frac{\partial}{\partial z} \sigma(z) = \frac{\partial}{\partial z} \frac{1}{1 + e^{-z}} = -(1 + e^{-z})^{-2} \frac{\partial}{\partial z} (1 + e^{-z})$$

$$= \frac{e^{-z}}{(1 + e^{-z})^2} = \left(\frac{1}{1 + e^{-z}} \right) \left(\frac{e^{-z}}{1 + e^{-z}} \right) = \sigma(z) \left\{ \frac{1 + e^{-z}}{1 + e^{-z}} - \frac{1}{1 + e^{-z}} \right\}$$

$$= \sigma(z)(1 - \sigma(z))$$

Learning Parameters

Resuming from (1)

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n} \left[\frac{y_i}{\sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} \sigma_\theta(x_i) + \frac{1 - y_i}{1 - \sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} (1 - \sigma_\theta(x_i)) \right]$$

$$= -\sum_{i=1}^{n} \left[\frac{y_i \sigma_\theta(x_i)}{\sigma_\theta(x_i)} (1 - \sigma_\theta(x_i)) \frac{\partial}{\partial \theta_j} (x_i \theta) + \frac{1 - y_i}{1 - \sigma_\theta(x_i)} (1 - \sigma_\theta(x_i)) \frac{\partial}{\partial \theta_j} (1 - \sigma_\theta(x_i)) \right]$$

$$= -\sum_{i=1}^{n} \left[y_i (1 - \sigma_\theta(x_i)) x_i^j - (1 - y_i) \sigma_\theta(x_i) x_i^j \right]$$

$$= -\sum_{i=1}^{n} \left[(y_i - y_i \sigma_\theta(x_i) - \sigma_\theta(x_i) + y_i \sigma_\theta(x_i)) x_i^j \right]$$

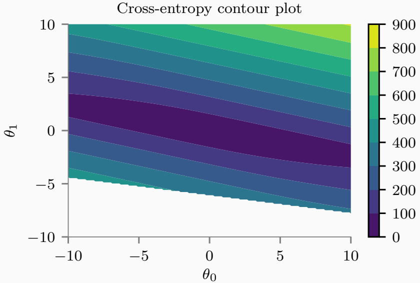
$$= \sum_{i=1}^{n} \left[\sigma_\theta(x_i) - y_i \right] x_i^j$$

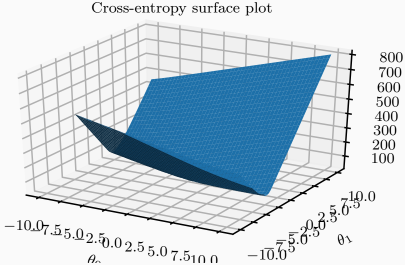
Learning Parameters

$$\boxed{\frac{\partial J(\theta)}{\theta_j} = \sum_{i=1}^{N} \left[\sigma_\theta(x_i) - y_i\right] x_i^j}$$

Now, just use Gradient Descent!

Cost function convexity





Hessian Matrix

The Hessian matrix of f(.) with respect to θ, written $\nabla^2_\theta f(\theta)$ or simply as $\mathbb{H}$, is the $d \times d$ matrix of partial derivatives,

$$\nabla^2_\theta f(\theta) = \begin{bmatrix} \frac{\partial^2 f(\theta)}{\partial \theta_1^2} & \frac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 f(\theta)}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 f(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f(\theta)}{\partial \theta_2^2} & \dots & \frac{\partial^2 f(\theta)}{\partial \theta_2 \partial \theta_n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 f(\theta)}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 f(\theta)}{\partial \theta_n \partial \theta_2} & \dots & \frac{\partial^2 f(\theta)}{\partial \theta_n^2} \end{bmatrix}$$

Newton's Algorithm

The most basic second-order optimization algorithm is Newton's algorithm, which consists of updates of the form,

$$\theta_{k+1} = \theta_k - \mathbb{H}_k^{1} g_k$$

where g_k is the gradient at step k. This algorithm is derived by making a second-order Taylor series approximation of $f(\theta)$ around θ_k:

$$f_{quad}(\theta) = f(\theta_k) + g_k^T(\theta - \theta_k) + \frac{1}{2}(\theta - \theta_k)^T \mathbb{H}_k(\theta - \theta_k)$$

differentiating and equating to zero to solve for θ_{k+1}.

Learning Parameters

Now assume:

$$g(\theta) = \sum_{i=1}^{n} \Big[\sigma_\theta(x_i) - y_i\Big] x_i^j = \mathsf{X}^\mathsf{T}(\sigma_\theta(\mathsf{X}) - \mathsf{y})$$

$$\pi_i = \sigma_\theta(x_i)$$

Let $\mathbb{H}$ represent the Hessian of $J(\theta)$

$$\begin{aligned}
\mathbb{H} = \frac{\partial}{\partial\theta} g(\theta) &= \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \Big[\sigma_\theta(x_i) - y_i\Big] x_i^j \\
&= \sum_{i=1}^{n} \left[\frac{\partial}{\partial\theta}\sigma_\theta(x_i) x_i^j - \frac{\partial}{\partial\theta} y_i x_i^j\right] \\
&= \sum_{i=1}^{n} \sigma_\theta(x_i)(1 - \sigma_\theta(x_i)) x_i x_i^T \\
&= \mathsf{X}^\mathsf{T} diag(\sigma_\theta(x_i)(1 - \sigma_\theta(x_i)))\mathsf{X}
\end{aligned}$$

Iterratively reweighted least squares (IRLS)

For binary logistic regression, recall that the gradient and Hessian of the negative log-likelihood are given by:

$$
\begin{aligned}
g(\theta)_k &= \mathsf{X}^{\mathsf{T}}(\pi_{\mathsf{k}} - \mathsf{y}) \\
\mathsf{H}_k &= \mathsf{X}^T S_k \mathsf{X} \\
\mathsf{S}_k &= diag(\pi_{1k}(1-\pi_{1k}), \ldots, \pi_{nk}(1-\pi_{nk})) \\
\pi_{ik} &= sigm(\mathsf{x}_{\mathsf{i}}\theta_{\mathsf{k}})
\end{aligned}
$$

The Newton update at iteraion $k + 1$ for this model is as follows:

$$
\begin{aligned}
\theta_{k+1} &= \theta_k - \mathbb{H}^{-1} g_k \\
&= \theta_k + (X^T S_k X)^{-1} X^T (y - \pi_k) \\
&= (X^T S_k X)^{-1} [(X^T S_k X)\theta_k + X^T (y - \pi_k)] \\
&= (X^T S_k X)^{-1} X^T [S_k X \theta_k + y - \pi_k]
\end{aligned}
$$

Regularized Logistic Regression

Unregularised:

$$J_1(\theta) = -\left\{ \sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1 - y_i) \log(1 - \sigma_\theta(x_i)) \right\}$$

L2 Regularization:

$$J(\theta) = J_1(\theta) + \lambda \theta^T \theta$$

L1 Regularization:

$$J(\theta) = J_1(\theta) + \lambda |\theta|$$

Multi-Class Prediction

1. Use one-vs.-all on Binary Logistic Regression
2. Use one-vs.-one on Binary Logistic Regression
3. Extend Binary Logistic Regression to Multi-Class Logistic Regression

Softmax

$$Z \in \mathbb{R}^d$$

$$f(z_i) = \frac{e^{z_i}}{\sum_{i=1}^{d} e^{z_i}}$$

$$\therefore \sum f(z_i) = 1$$

$f(z_i)$ refers to probability of class i

Softmax for Multi-Class Logistic Regression

$$k = 1, \ldots, k \text{classes}$$

$$\boxed{P(y = k|x, \theta) = \frac{e^{x\theta_k}}{\sum_{k=1}^{K} e^{x\theta_k}}}$$

Softmax for Multi-Class Logistic Regression

For K = 2 classes,

$$P(y = k|x, \theta) = \frac{e^{x\theta_k}}{\sum_{k=1}^{K} e^{x\theta_k}}$$

$$P(y = 0|x, \theta) = \frac{e^{x\theta_0}}{e^{x\theta_0} + e^{x\theta_1}}$$

$$P(y = 1|x, \theta) = \frac{ex\theta_1}{e^{x\theta_0} + e^{x\theta_1}} = \frac{e^{x\theta_1}}{e^{x\theta_1}\{1 + e^{x(\theta_0 - \theta_1)}\}}$$

$$= \frac{1}{1 + e^{-x\theta'}}$$

$$= \text{Sigmoid!}$$

Multi-Class Logistic Regression Cost

For 2 class we had:

$$J(\theta) = -\left\{ \sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1 - y_i) \log(1 - \sigma_\theta(x_i)) \right\}$$

Extend to K-class:

$$J(\theta) = \left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} I\{y_i = k\} \log \frac{e^{x_i \theta_k}}{\sum_{k=1}^{K} e^{x_i \theta_k}} \right\}$$

$i \rightarrow$ Sample #

$k \rightarrow Class$

I: Identity Function

I(true) = 1; I(false) = 0

Multi-Class Logistic Regression Cost

Now:

$$\frac{\partial J(\theta)}{\partial \theta_k} = \sum_{i=1}^{n} \left[x_i \left\{ I(y_i = k) - P(y_i = k | x_i, \theta) \right\} \right]$$